Information Systems

Introduction

Big Data Analytics

Presented by: Dr Sherin El Gokhy

Basic Methods

Module 3 – Review of Basic Data Analytic Methods Using R

Module 3: Review of Basic Data Analytic Methods Using R

Upon completion of this module, you should be able to:

- Use basic analytics methods such as distributions, statistical tests and summary operations to investigate a data set.
- Use R as a tool to perform basic data analytics, reporting and basic data visualization.

Putting the Data Analytics Lifecycle into Practice

- From Module 2 you learned a strategy to approach any data analytics problem:
    - **Phase 1: Discovery**
    - **Phase 2: Data Preparation**
    - **Phase 3: Model Planning** *(covered in Module 4)*
    - Phase 4: Model Building
    - Phase 5: Communicate Results
    - Phase 6: Operationalize

- To begin to analyze the data you need:
    - 1. A tool that allows you to look at the data – that is "R".
    - 2. Skill in basic statistics – we're providing a refresher.

Module 3: Review of Basic Data Analytic Methods Using R

Part 1: Using R to Look at Data – Introduction to R

During this lesson the following topics are covered:

- Using the R Graphical User Interface
- Overview: Getting Data into (and out of) R
- Data Types Used in R
- Basic R Operations
- Basic Statistics
- Generic Functions

GETTING A HANDLE ON THE DATA

Five Things to Remember About R

1. (Almost) everything is a *object*

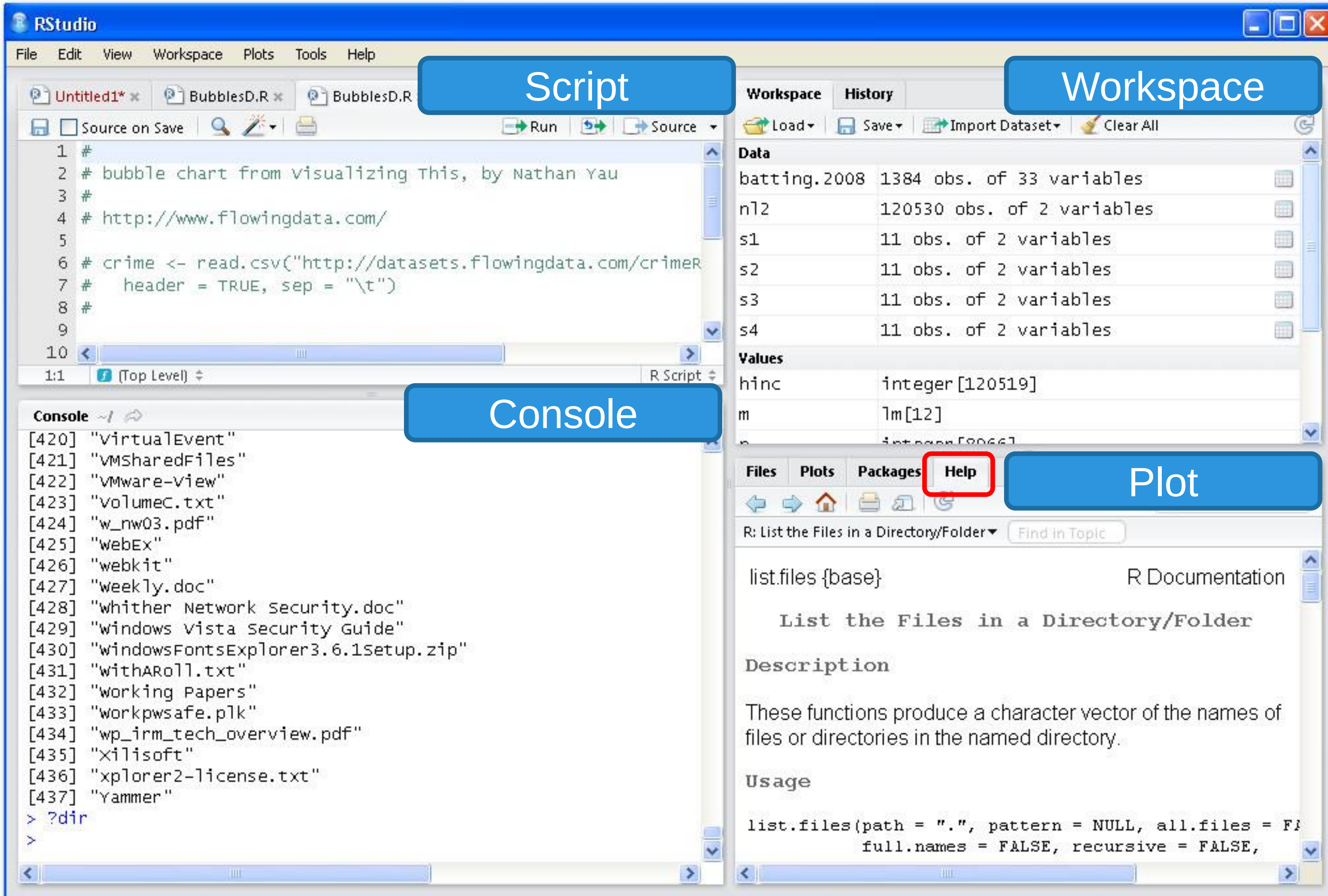
2. (Almost) everything is a *vector*
   - Example: `x <- 3` -- `x` is a vector of length 1
     `v <- c(2,4,6,8,10)` -- `v` is a vector of length 5

3. All commands are functions
   - Example: `quit()` or `q()`, not q

4. Some commands produce different output depending…

5. Know your default arguments!

Using the RStudio Graphical User Interface

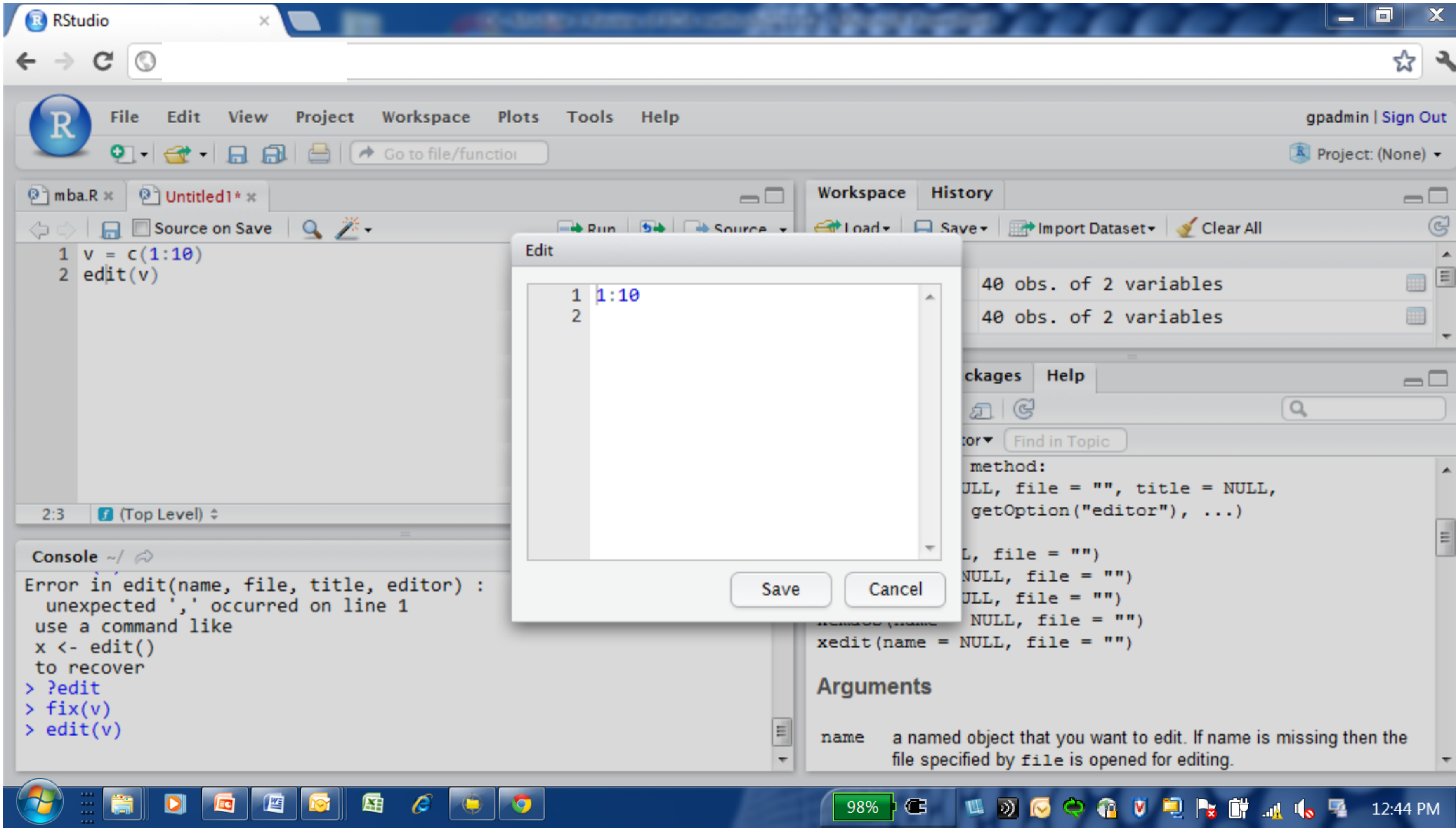


Overview: Getting Data Into (and Out of) R

- Getting Data Into R
  - Type it in (if it's small)!
  - Read from a data file
  - Read from a database

- Getting Data Out of R
  - Save in a workspace
  - Write a text file
  - Save an object to the file system
  - You can save plots as well!

Typing Data Into R



Getting Data Into R: External Sources

- R supports multiple file formats
  - `read.table()` is the main function
- File name can be a URL
  - `read.table("http://ahost/file.csv", sep=",")` is the same as `read.csv(…)`
- Can read directly from a database via ODBC interface
  - `mydb <- odbcConnect("MyPostgresDB", …)`
- R packages exist to read data from Hadoop or HDFS (more later)

Note! R always uses the forward-slash ("/") character in full file names
"C:/users/janedoe/My Documents/Script.R"

Getting Data Out of R

Options	R Code

Save it as part of your workspace (or a different workspace)	`save.image(file="dfm.Rdata")`
	`save.image()  # a .Rdata file`
	`load.image("dfm.Rdata")`

Save it as a data file	`write.csv(dfm, file="dfm.csv")`

Save it as an R object	`save(Mydata,`
	`     file="Mydata.Rdata")`
	`load(file="Mydata.Rdata")`

Plots can be saved as images	`saveplot(filename="`*`filename.ext`*`",` `    type="`*`type`*`")`

Data Classification: A Quick Review

Data “Noir”	Examples
Nominal	condo, house, rental
Ordinal	hates < dislikes <neutral < likes < loves
Interval	10F colder tomorrow than today
Ratio	5342 > 4321

Some statistical tests require data at the interval level or higher. Other tests assume ordinal or nominal. Make sure you check.

Data Types Used in R

Data Types	R Code
Numbers, Strings	n <- 3 s <- “columbus, ohio”
Vectors	levels <- c(“Wow”, “Good”,“Bad”) ratings <- c(“Bad”, “Bad”, “Wow”)
Factors and Lists	f <- factor(ratings, levels) l <- list(ratings=ratings, critics=c(“Siskel”,“Ebert”))
Functions	stdev <- function(x) {sd(x)}

R Structured Types

Data Types	R Code
Matrix - (n*m numeric data frame)	m <- matrix(c(1:3, 11:13), nrow = 2, ncol = 3, byrow = TRUE)
Table – contingency table	t <- table(dfm$factor_variable)
Data frames – data sets	dfm <- read.csv(“CrimeRatesByStates2005.csv”)
Extracting data	xdfm <- dfm[1:3,] ydfm <- dfm[, 3:5] s <- dfm$state

Basic R Operations on Vectors

Function	R Code
Operations on Vectors	v <- c(1:10); w <- c(15:24) ; nv <- v * pi ; nw <- w * v
Vector transformations	radius <- sqrt(d$area)/ pi) t <- as.table(dfm$factor_variable) pct <- t/sum(t)* 100
Logical Vectors	v[v < 1000] ndf <- subset(dfm, d$population < 10000) nv <- v[c(1,2,3,5,8,13)]
Examining data structures	dim(dfm); attributes(dfm) ; class(dfm); typeof(dfm)

Descriptive Statistics

Function	R Code
View the data	head(x); tail(x)
View a summary of the data	summary(x)
Compute basic statistics	sd(x); var(x); range(x); IQR(x)
Correlation	cor(x); cor(d$var1, d$var2)

Generic Functions

- Also known as method overriding in OO-land

- Specific actions that differ based on the class of the object :

Code	Function
Plot the variable x	plot (x)
Histogram of x	hist (x)
Internal structure of x	str (x)

- Good for initial data exploration (more later)

Check Your Knowledge

- Which data structures in R are the most used? Why?
- Consider the cbind() function and the rbind() function that bind a vector to a data frame as a new column or a new row. When might these functions be useful?

Module 3: Review of Basic Data Analytic Methods Using R

Part 1: Summary

During this lesson the following topics were covered:

- How to use the R Graphical User Interface
- How to get data into (and out of) R
- Data Types used in R, and the basic R operations
- Basic descriptive statistics
- Using generic functions

Lab Exercise 2: Introduction to R

This lab is designed to investigate and practice working with R and using it to examine data.

- After completing the tasks in this lab you should able to:
  - Read data sets into R, save them, and examine the contents

Lab Exercise 2: Introduction to R

1 • Invoke the R environment

2 • Examine the Workspace

3 • Getting Familiar with R

4 • Read in the Lab Script

5 • Working with R : reading external data

6 • Verify the contents of the tables

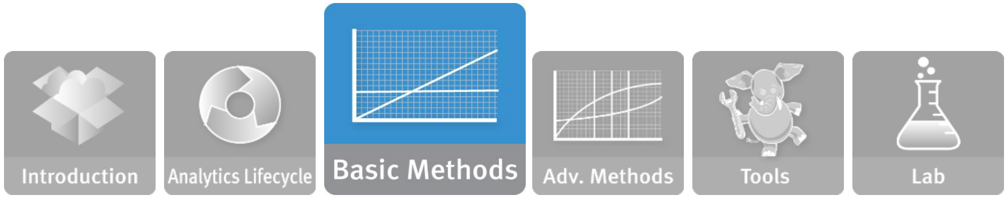
7 • Manipulating data frames in R

8 • Investigate your data

9 • Save the data sets

10 • Continue investigating the data

11 • Exit R



Module 3: Review of Basic Data Analytic Methods Using R

Part 2: Analyzing and Exploring the Data

During this lesson the following topics are covered:

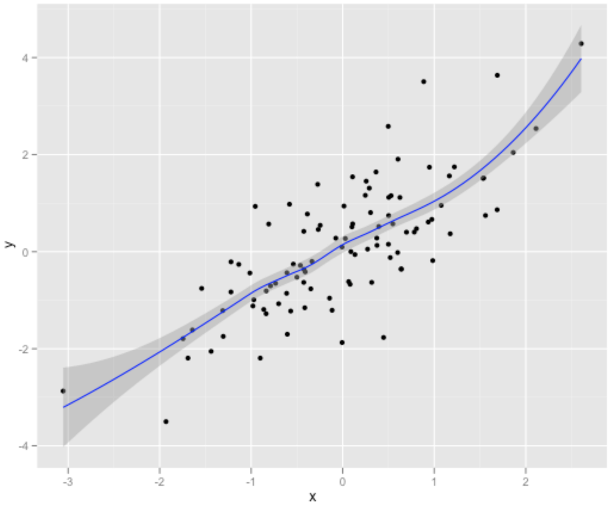
- Why visualize?
- Examining a single variable
- Examining pairs of variables
- Indications of dirty data.
- Data exploration vs. presentation

Why Visualize?

Summary statistics give us some sense of the data:

- Mean vs. Median.
- Standard deviation.
- Quartiles, Min/Max.
- Correlations between variables.

```
summary(data)
       x                  y
 Min.   :-3.05439   Min.   :-3.50179
 1st Qu.:-0.61055   1st Qu.:-0.75968
 Median : 0.04666   Median : 0.07340
 Mean   :-0.01105   Mean   : 0.09383
 3rd Qu.: 0.56067   3rd Qu.: 0.88114
 Max.   : 2.60614   Max.   : 4.28693
```



Visualization gives us a more holistic sense

Anscombe's Quartet

4 data sets, characterized by the following. Are they the same, or are they different?

Property	Values
Mean of x in each case	9
Exact variance of x in each case	11
Exact mean of y in each case	7.5 (to 2 d.p)
Variance of Y in each case	4.13 (to 2 d.p)
Correlations between x and y in each case	0.816
Linear regression line in each case	Y = 3.00 + 0.500x (to 2 d.p and 3 d.p resp.)

i

X	y
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.84
7.00	4.82
5.00	5.68

ii

x	y
10.00	9.14
8.00	8.14
13.00	8.74
9.00	8.77
11.00	9.26
14.00	8.10
6.00	6.13
4.00	3.10
12.00	9.13
7.00	7.26
5.00	4.74

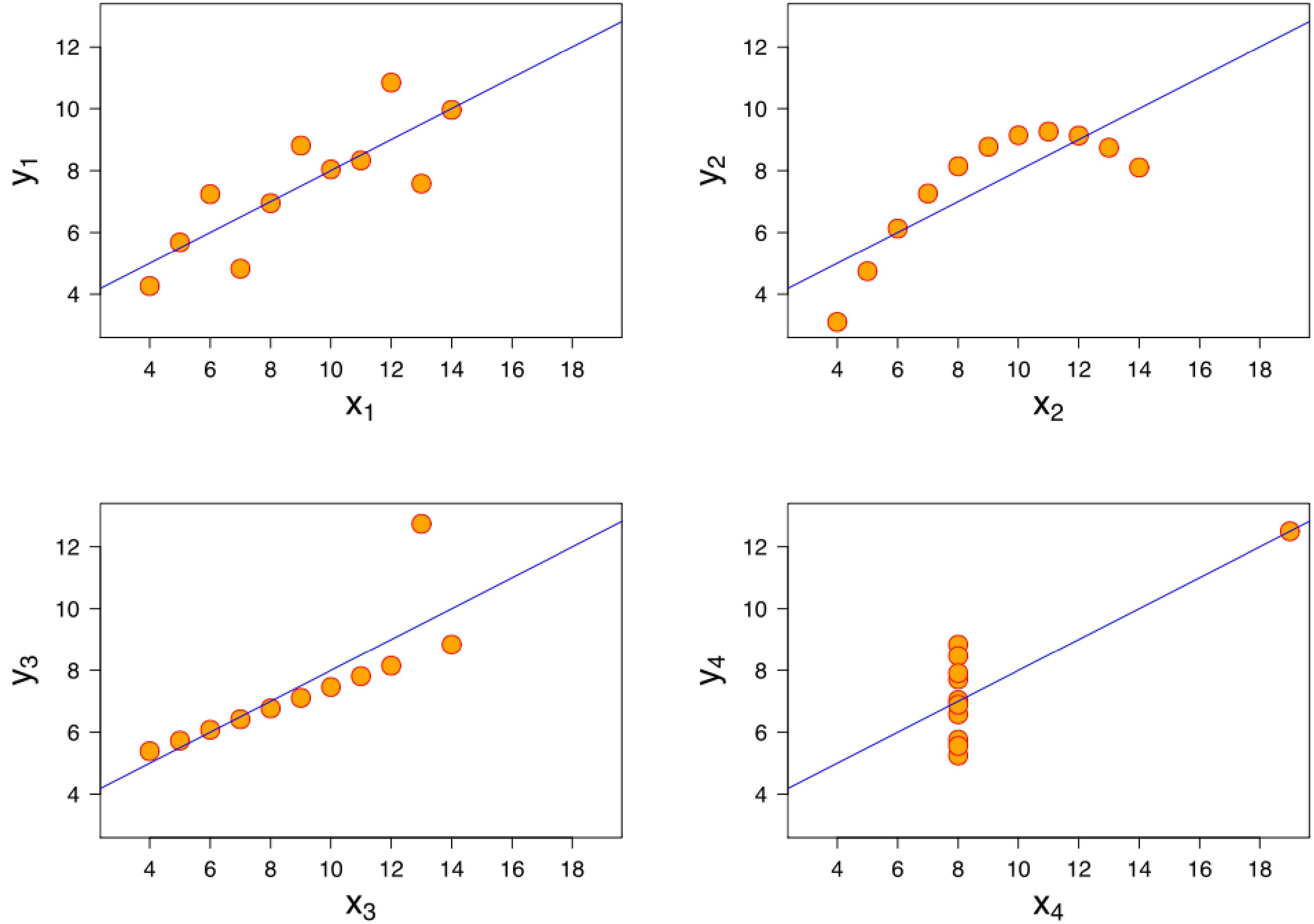
iii

x	y
10.00	7.46
8.00	6.77
13.00	12.74
9.00	7.11
11.00	7.81
14.00	8.84
6.00	6.08
4.00	5.39
12.00	8.15
7.00	6.42
5.00	5.73

iv

x	y
8.00	6.58
8.00	5.76
8.00	7.71
8.00	8.84
8.00	8.47
8.00	7.04
8.00	5.25
19.00	12.50
8.00	5.56
8.00	7.91
8.00	6.89

Moral: Visualize Before Analyzing!



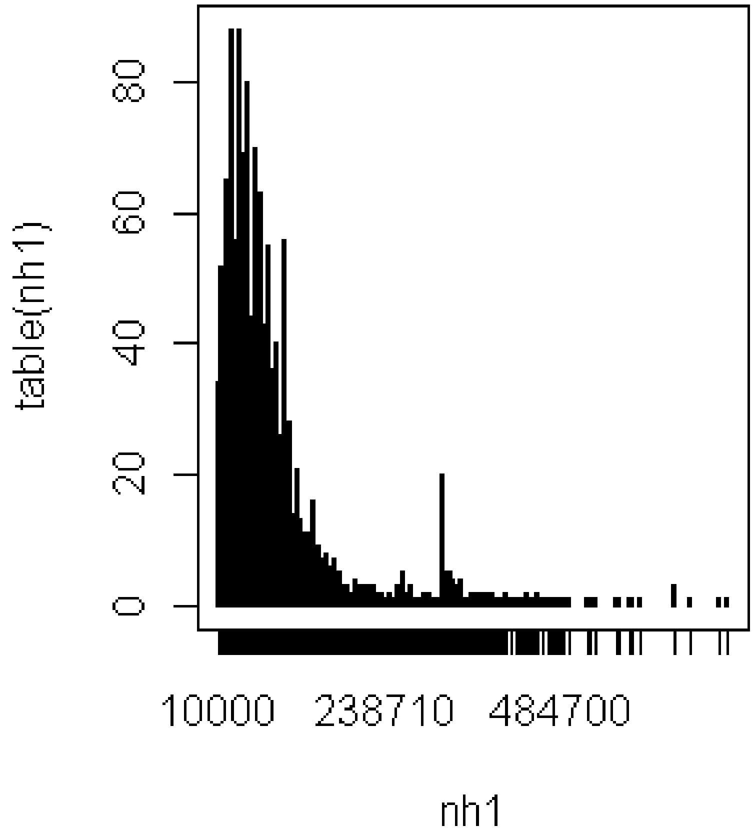
Visualizing Your Data

- Examining the distribution of a single variable
- Analyzing the relationship between two variables
- Establishing multiple pair wise relationships between variables
- Analyzing a single variable over time
- Data exploration versus data presentation

Examining the Distribution of a Single Variable

Graphing a single variable

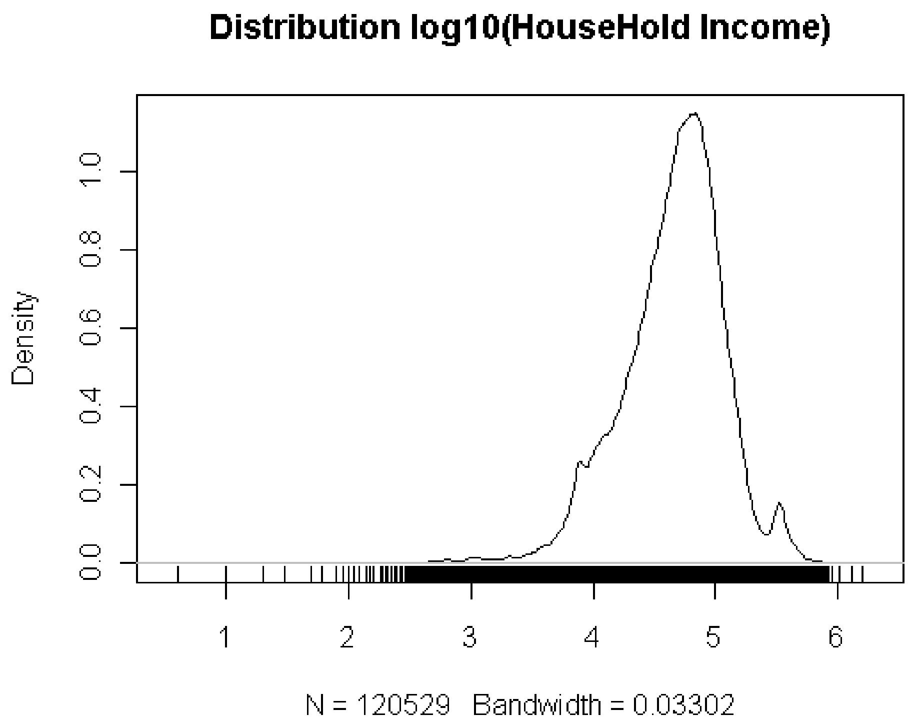
- plot(sort(.)) – for low volume data
- hist(.) – a histogram
- plot(density(.)) – densityplot
  - A "continuous histogram“
- Example
  - Frequency table of household income



Examining the Distribution of a Single Variable

Graphing a single variable

- plot(sort(.)) – for low volume dá
- hist(.) – a histogram
- plot(density(.)) – densityplot
  - A "continuous histogram“
- Example
  - Frequency table of household income
    - rug() plot emphasizes distribution



What are we looking for?

A sense of the data range

- If it's very wide, or very skewed, try computing the log

Outliers, anomalies

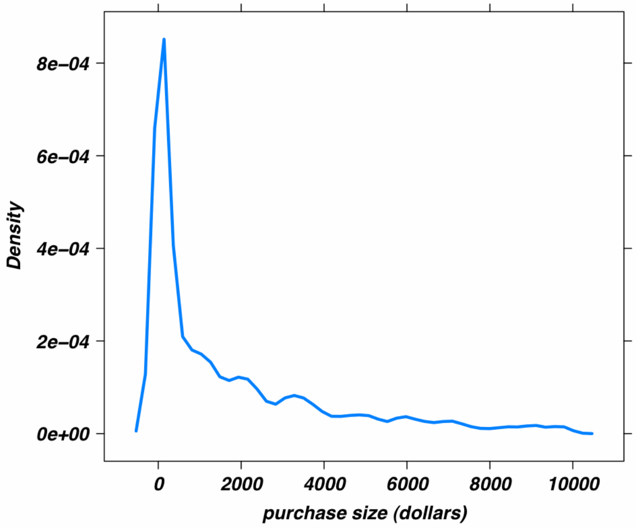
- Possibly evidence of dirty data

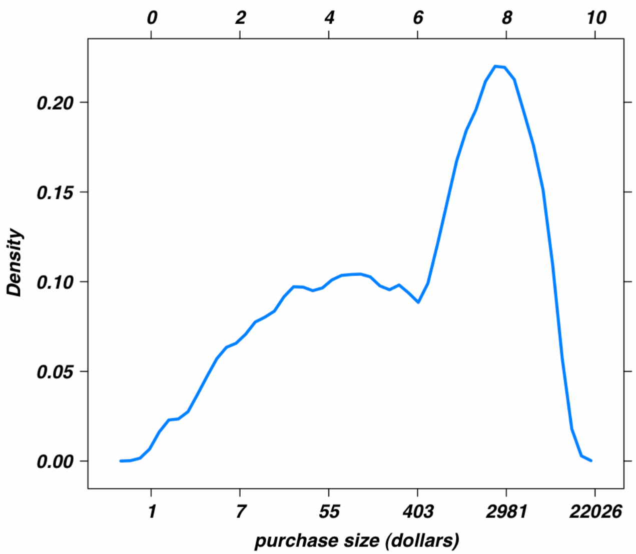
Shape of the Distribution

- Unimodal? Bimodal?
- Skewed to left or right?
- Approximately normal? Approximately lognormal?

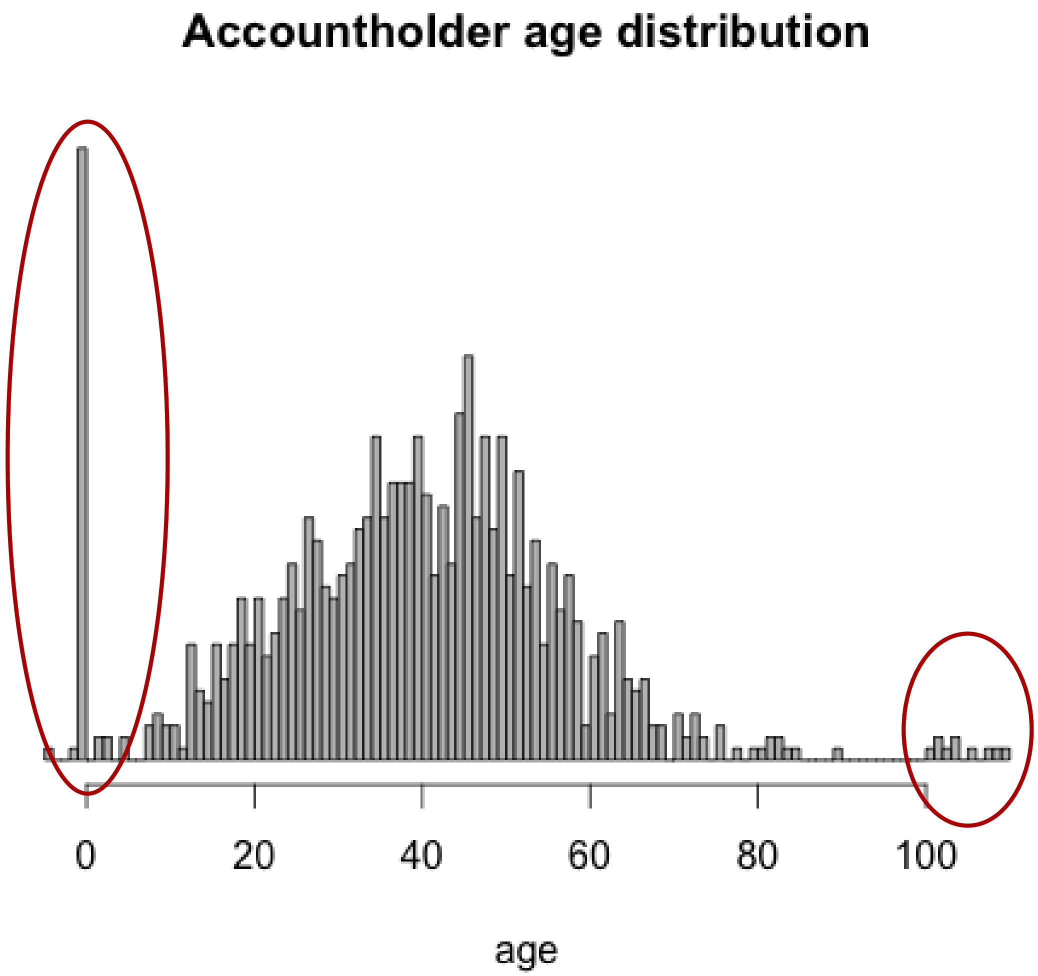
Example - Distribution of purchase size ($)

- Range from 0 to > $10K, right skewed
- Typical of monetary data
- Plotting log of data gives better sense of distribution
- Two purchasing distributions
  - ~ $55
  - ~ $2900





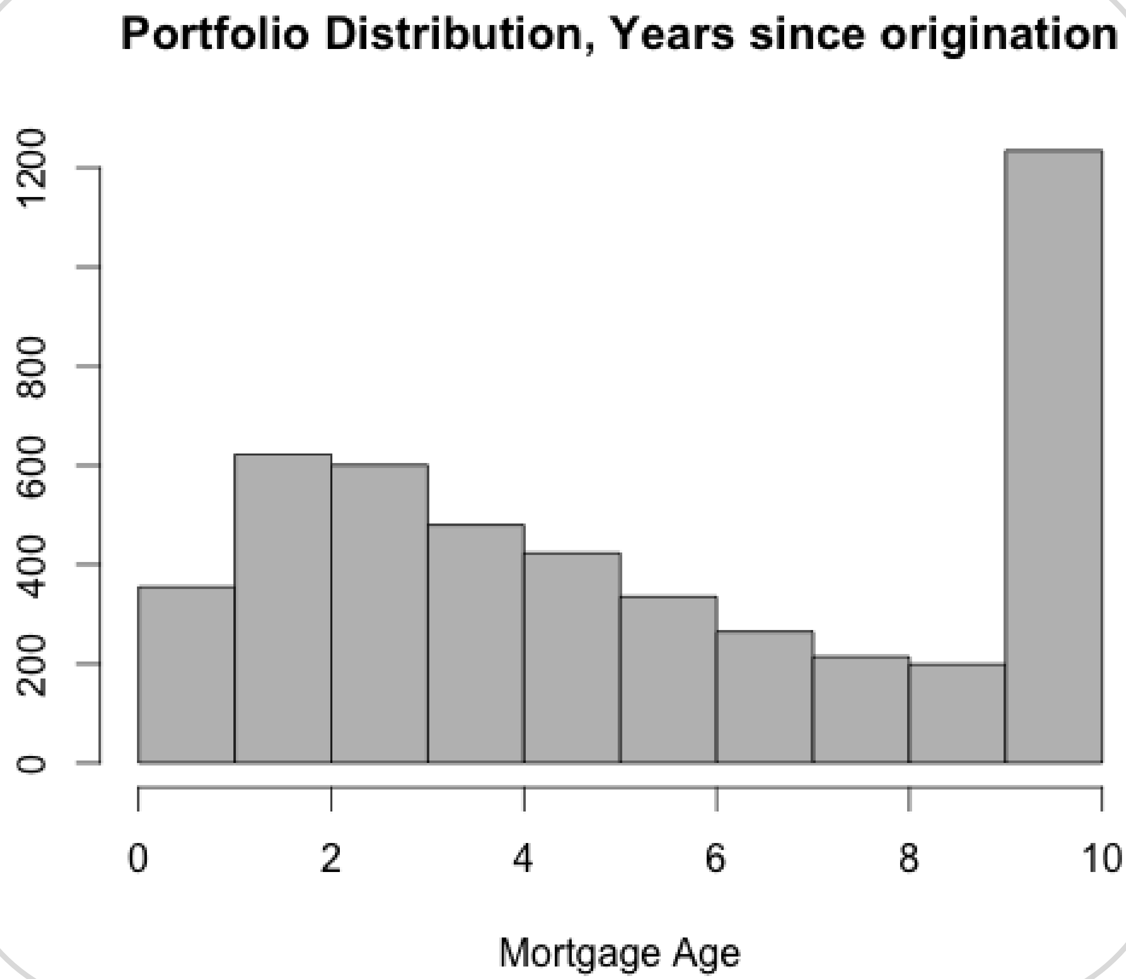
Evidence of Dirty Data



Missing values?

Mis-entered data? Inherited accounts?

"Saturated" Data



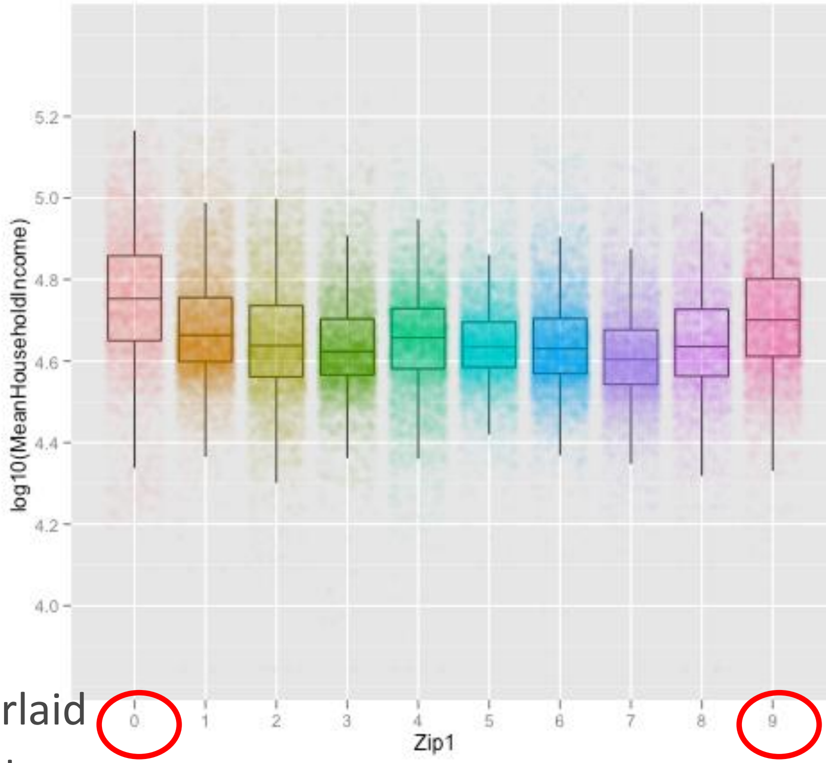
Do we really have no mortgages older than 10 years?

Or does the year 2004 in the origination field mean "2004 or prior"?

Analyzing the Relationship Between Two Variables

How?

- Two Continuous Variables (or two discrete variables)
  - Scatterplots
  - LOESS (fit smoothed line to the data)
  - Linear models: graph the correlation
  - Binplots, hexbin plots
    - More legible color-based plots for high volume data
- Continuous *vs.* Discrete Variable
  - Jitter, Box and whisker plots, Dotplot or barchart

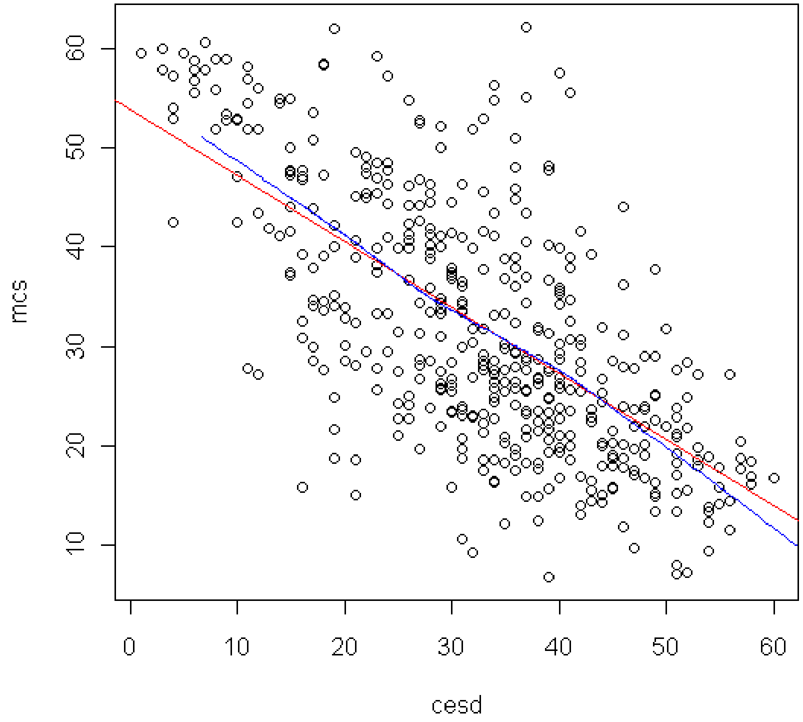


Example:

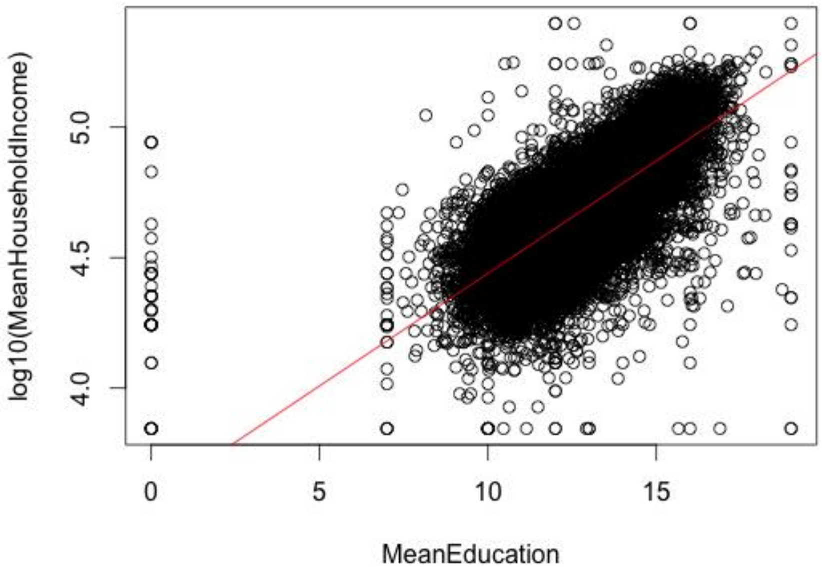
- Household income by region (ZIP1)
- Scatterplot with jitter, with box-and-whisker overlaid
- New England (0) and West Coast (9) have highest mean household income

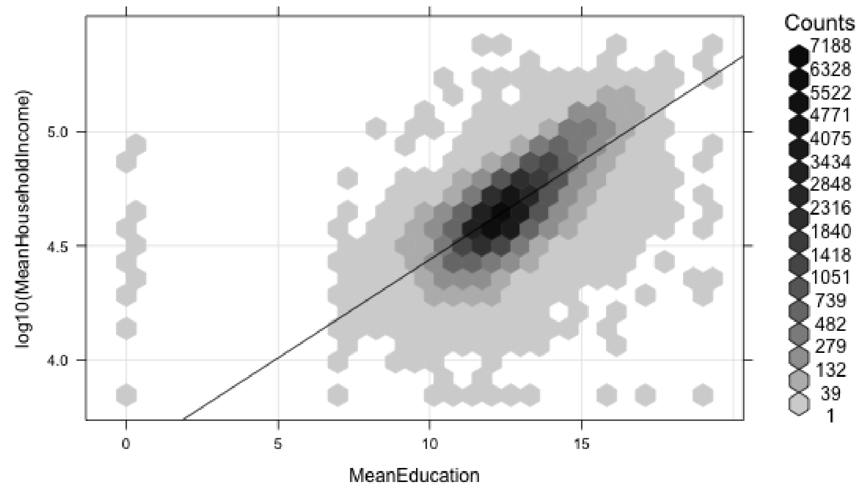
Two Variables: What are we looking for?

- Is there a relationship between the two variables?
  - Linear? Quadratic?
  - Exponential?
    - Try semi-log or log-log plots
  - Is it a cloud?
    - Round? Concentrated? Multiple Clusters?
- How?
  - Scatterplots
- Example
  - Red line: linear fit
  - Blue line: LOESS
  - Fairly linear relationship, but with wide variance



Two Variables: High Volume Data - Plotting



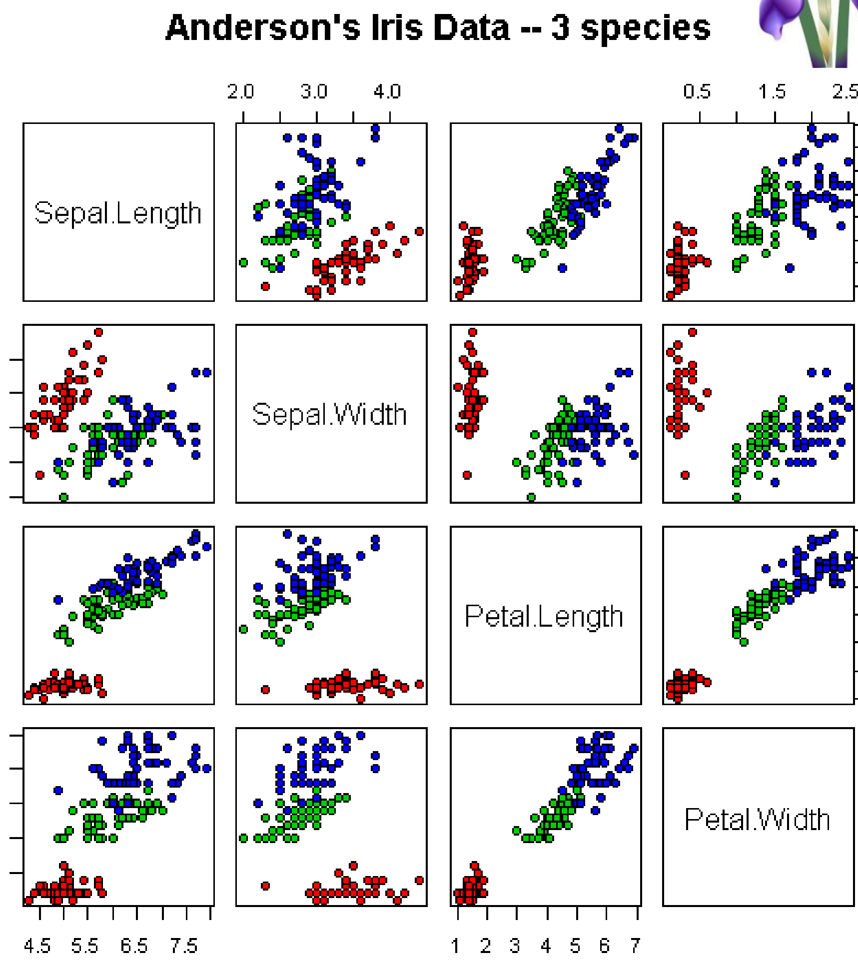


Scatterplot:
Overplotting makes it difficult to see structure

Hexbinplot:
Now we see where the data is concentrated.

Establishing Multiple Pairwise Relationships Between Variables

- Why?
  - Examine many two-way relationships quickly
- How?
  - pairs(ds) can generate a plot of each pairs of variables
- Example
  - Iris Characteristics
    - Strong linear relationship between petal length and width
    - Petal dimensions discriminate species more strongly than sepal dimensions



Analyzing a Single Variable over Time

What?

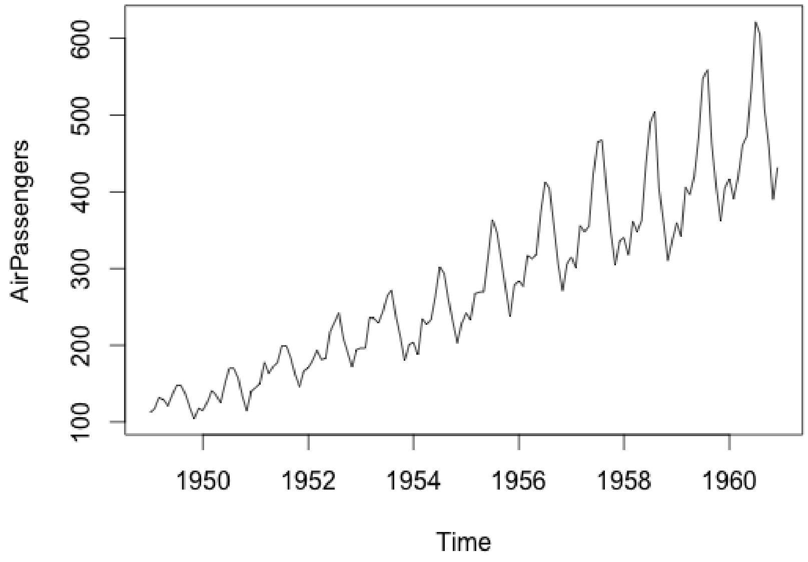
- Looking for …
  - Data range
  - Trends
  - Seasonality

How?

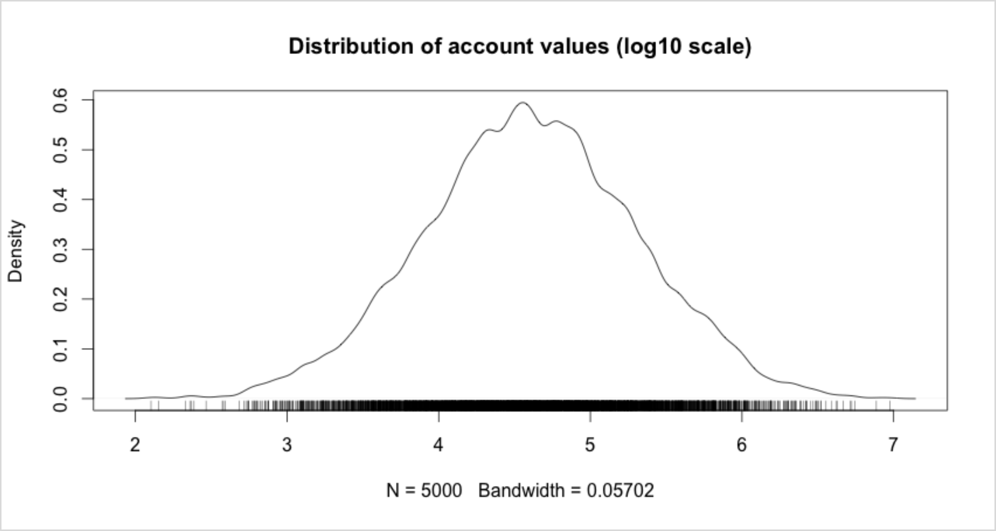
- Use time series plot

Example

- International air travel (1949-1960)
- Upward trend: growth appears superlinear
- Seasonality
  - Peak air travel around Nov. with smaller peaks near Mar. and June

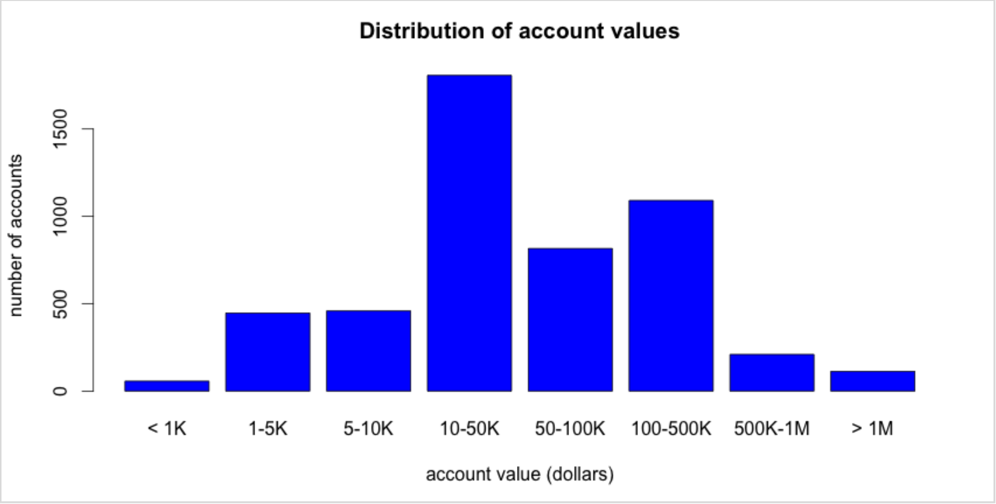


Data Exploration vs. Presentation



Data Exploration:

This tells you what you need to know.



Presentation:

This tells the stakeholders what they need to know.

Check Your Knowledge

- Do you think the regression line sufficiently captures the relationship between the two variables? What might you do differently?

- In the Iris slide example, how would you characterize the relationship between sepal width and sepal length?

- Did you notice the use of color in the Iris slide? Was it effective? Why or why not?

Module 3: Review of Basic Data Analytic Methods Using R

Part 2: Summary

During this lesson the following topics were covered:

- Justifying why we visualize data
- Using plots and graphs to determine:
  - Shape of a single variable
  - “dirty” data or “saturated” data
  - Relationship between two or more variables
  - Relationship between multiple variables
  - A single variable over time
- Data exploration *versus* Presentation

Thanks